ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

ROOM-3D: Real-Time Unsupervised Online 3D Room Segmentation

Rafael Flor-Rodríguez-Rabadán[1], Carlos Gutiérrez-Álvarez[1], Alexis Bañuls-González[1], Sergio Lafuente-Arroyo[1], Saturnino Maldonado-Bascón[1] and Roberto Javier López-Sastre[1]

***Abstract*— Room-level understanding is essential for mobile robots operating in indoor environments. Existing room segmentation methods assume an offline setting—requiring a complete scene reconstruction before producing the final result—which limits their applicability to real-time robotic navigation. We introduce the novel problem of *online 3D room segmentation*, where a robot must continuously segment rooms and detect transitions from streaming observations during exploration, and propose ROOM-3D: a real-time unsupervised framework that combines Gaussian-based SLAM with open-vocabulary semantic reasoning to incrementally build a 3D room segmentation without access to future observations or global post-processing. We also introduce instantaneous evaluation metrics tailored to this online setting. Experiments on HM3D-Semantics demonstrate temporally consistent, accurate segmentation under strict online constraints, with state-of-the-art results in the offline evaluation too.**

I. INTRODUCTION AND MOTIVATION

The objective of room segmentation is to infer a coherent spatial model of an indoor environment, generating a structured map of its constituent rooms. Most existing approaches operate *offline*—from geometry-based pipelines [1], [2] to learning-based models such as RoomFormer [3] or HOV-SG [4]—and do not support incremental understanding. LEXIS [5] augments a topological SLAM graph with CLIP features but, unlike ROOM-3D, does not produce an explicit geometric 3D segmentation updated during navigation. Our goal is to segment rooms *incrementally as the robot navigates*—the novel *online 3D room segmentation* problem—enabling real-time applications such as service robots, inspection drones, or assistive systems in complex indoor spaces.

Contributions. (i) We formalize the novel *online 3D room segmentation* problem. (ii) We propose ROOM-3D, combining Gaussian SLAM with CLIP-based open-vocabulary reasoning. (iii) We introduce *instantaneous evaluation metrics* for online room segmentation and transition detection. (iv) Experiments on HM3D-Semantics show state-of-the-art performance in both online and offline settings.

II. METHOD OVERVIEW

ROOM-3D constructs an online 3D room segmentation map from a stream of RGB-D images using three tightly integrated components (see Fig. 1).

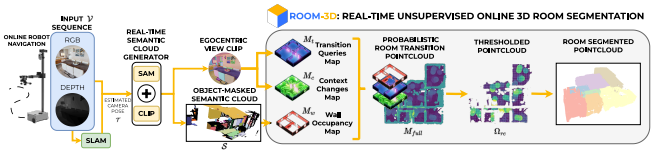


Fig. 1: ROOM-3D pipeline. From streaming RGB-D input, the system builds an open-vocabulary 3D map via Gaussian SLAM, SAM, and CLIP. Room segmentation fuses wall occupancy (M_w), transition (M_t), and context change (M_c) maps to extract room boundaries and detect transitions in real time.

Open-Vocabulary 3D Mapping. Building upon OVO [6], the module integrates a Gaussian-SLAM [7] backbone that incrementally reconstructs the scene from posed RGB-D keyframes. SAM [8] produces class-agnostic 2D segments, which are associated with existing 3D segments or used to initialize new ones. Each 3D segment is represented by a CLIP embedding, enabling open-vocabulary semantic queries via cosine similarity.

Room Segmentation Model. The segmentation integrates three complementary cue maps, each encoding boundary evidence as $M : \mathbb{R}^3 \to [0, 1]$:

- *Wall occupancy map* (M_w): identifies room-bounding structures via CLIP similarity against queries such as "wall" or "doorway".
- *Transition map* (M_t): detects navigable passages via zero-shot CLIP comparison against transition-related queries, projected onto floor points for temporal robustness.
- *Context map* (M_c): captures semantic shifts by clustering egocentric CLIP embeddings with incremental MiniBatchKMeans, handling open-plan layouts without geometric boundaries.

These maps are fused into $M_{\text{full}} = M_w + M_t + M_c$. Low-probability regions (below threshold th) indicate room interiors; connected components within these regions define individual rooms with unique labels. Segmentation updates incrementally as the map evolves.

Live Room Transition Detection. A mapping function L assigns to each keyframe pose its enclosing room label; a transition is flagged whenever consecutive labels differ, yielding a binary per-frame signal with no additional detector.

Online Evaluation Metrics. We introduce *instantaneous precision* (iPrec) and *instantaneous recall* (iRec) for room segmentation, and analogous metrics (iTrPrec/iTrRec) for transition detection, evaluated at reg-

[1]All authors are with the Department of Signal Theory and Communications, University of Alcalá, Alcalá de Henares, Madrid, Spain. rafael.flor@uah.es

TABLE I: Overall results on 8 HM3D-Semantics trajectories. Top: online evaluation (ours). Bottom: offline comparison (Prec/Rec [%]).

Online Evaluation (ours)						
	m-iP	**m-iR**	**m-iF1**	**m-iTrP**	**m-iTrR**	**m-iTrF1**
ROOM-3D	**81.78**	**90.71**	**85.44**	**66.02**	**55.33**	**56.76**

Offline Comparison						
	Hydra [1]		HOV-SG [4]		ROOM-3D (ours)	
	Prec	Rec	Prec	Rec	Prec	Rec
Overall	**86.18**	77.55	84.10	83.59	85.75	**89.17**

ular intervals t_k and averaged across scenes. Given estimated rooms $R_e^{(t_k)}$ and ground-truth rooms $R_g^{(t_k)}$ over the observed region at t_k:

$$\begin{aligned} \text{iPrec}(t_k) &= \frac{1}{|R_e^{(t_k)}|} \sum_{r_e \in R_e^{(t_k)}} \max_{r_g \in R_g^{(t_k)}} \frac{|r_g \cap r_e|}{|r_e|}, \\ \text{iRec}(t_k) &= \frac{1}{|R_g^{(t_k)}|} \sum_{r_g \in R_g^{(t_k)}} \max_{r_e \in R_e^{(t_k)}} \frac{|r_e \cap r_g|}{|r_g|}. \end{aligned} \tag{1}$$

Sequence-level m-iPrec and m-iRec are averaged over all steps, and dataset-level m-iF1 captures both over- and under-segmentation across the full trajectory.

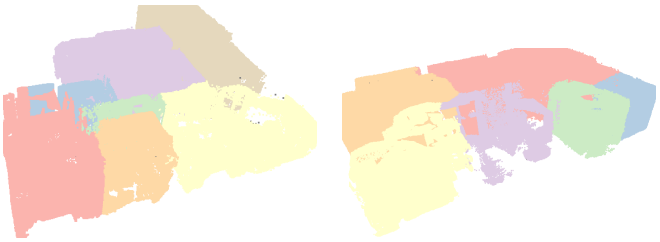
III. EXPERIMENTAL RESULTS

We evaluate on 8 trajectories from HM3D-Semantics [9], following the same scenes used by prior offline works [1], [4]. The selected trajectories span diverse architectural complexity, including open-plan layouts where adjacent regions share continuous space without geometric boundaries, and multi-floor buildings with up to 36 rooms.

Overall Performance. Table I summarizes the overall results of ROOM-3D (th=0.8, r_{wall}=0.3 m) averaged across all 8 trajectories, and compares against Hydra [1] and HOV-SG [4] using their standard offline metrics. Under our proposed online protocol, ROOM-3D achieves m-iF1 = 85.44% and m-iTrF1 = 56.76%, with peak per-scene values of 92.02% and 82.17%. In the offline setting, ROOM-3D obtains the best recall (89.17%) with competitive precision (85.75%), outperforming baselines that suffer from systematic under-segmentation.

Ablation and Runtime. Removing M_c drops m-iTrF1 from 56.76% to 38.64%; removing M_t further reduces m-iF1 to 78.72%, confirming that all three cues are complementary and essential. The end-to-end pipeline operates at 1.5 FPS, with the SAM+CLIP mapping stage (1.2 FPS) as the main bottleneck.

Qualitative Results. Fig. 2 shows color-coded 3D reconstructions of two HM3D scenes; ROOM-3D correctly partitions complex layouts into distinct functional volumes, including open-plan and non-structural boundaries.



(a) Scene 00824. (b) Scene 00829.

Fig. 2: Qualitative 3D room segmentation by ROOM-3D on two HM3D-Semantics scenes. Each color represents a distinct room identified incrementally during online exploration.

IV. CONCLUSIONS

We introduced *online 3D room segmentation* as a novel problem and proposed ROOM-3D, a real-time unsupervised framework combining Gaussian SLAM with CLIP reasoning for incremental room segmentation and transition detection. New instantaneous metrics enable principled temporal assessment, and experiments on HM3D-Semantics confirm state-of-the-art performance even against offline baselines. Future work will target scalability via sub-mapping and hierarchical scene graph reasoning for language-grounded navigation.

REFERENCES

[1] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," *RSS*, 2022.

[2] R. Bormann, F. Jordan, W. Li, J. Hampp, and M. Hägele, "Room segmentation: Survey, implementation, and analysis," *ICRA*, 2016.

[3] Y. Yue, T. Kontogianni, K. Schindler, and F. Engelmann, "Connecting the dots: Floorplan reconstruction using two-level queries," in *CVPR*, 2023.

[4] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," *RSS*, 2024.

[5] C. Kassab, M. Mattamala, L. Zhang, and M. Fallon, "Language-extended indoor slam (lexis): A versatile system for real-time visual scene understanding," *ICRA*, 2024.

[6] T. B. Martins, M. R. Oswald, and J. Civera, "Open-vocabulary online semantic mapping for slam," *IEEE Robotics and Automation Letters*, 2024.

[7] S. Zhu, G. Wang, H. Blum, J. Liu, L. Song, M. Pollefeys, and H. Wang, "SNI-SLAM: Semantic Neural Implicit SLAM," in *CVPR*, 2024.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Doll, and R. Girshick, "Segment Anything," *arXiv*, 2023.

[9] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva *et al.*, "Habitat-matterport 3d semantics dataset," *arXiv*, 2022.